A STUDY OF A FIT INDEX FOR EXPLANATORY ITEM RESPONSE THEORY MODELS

A Thesis
Presented to
The Academic Faculty

by

Heather Handy

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Psychology in the
School of Psychology

Georgia Institute of Technology
December 2019

A STUDY OF A FIT INDEX FOR EXPLANATORY ITEM RESPONSE THEORY MODELS

Approved by:

Dr. Susan Embretson, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Rick Thomas
School of Psychology
Georgia Institute of Technology

Dr. Michael Hunter
School of Psychology
Georgia Institute of Technology

Date Approved: [October 21, 2019]

ACKNOWLEDGEMENTS

I would like to thank my friend and labmate Clifford Hauenstein IV for this help and support. Without his help, it would have taken much longer to understand some of the nuances of the code in which I was programming. I would also like to thank my graduate advisor, Dr. Susan Embretson, who has been supportive and incredibly patient since the beginning of this endeavor regardless of any setbacks that have occurred along the way. Finally, I would like to thank my late mother who always encouraged me to try my best in everything. So as long as I put in as much effort as I could, she was proud no matter the outcome, which helped push me to have as much success as I do today.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **iii**

LIST OF FIGURES **vi**

LIST OF SYMBOLS AND ABBREVIATIONS **vii**

SUMMARY **ix**

CHAPTER 1. Introduction **1**
1.1 Four Item Response Model Types **3**
1.1.1 Doubly Descriptive 4
1.1.2 Item Explanatory 4
1.1.3 Person Explanatory 4
1.1.4 Doubly Explanatory 5
1.2 Item Response Theory Models for Analysis **5**
1.2.1 Rasch Models 5
1.2.2 Linear Logistic Test Model 6
1.2.3 Fit Statistic Models 8

CHAPTER 2. Methods **11**
2.1 Data Simulation **11**
2.1.1 Data Conditions 12
2.1.2 Dependent Measures 13
2.1.3 Generating True Item Difficulties 14
2.1.4 Item Response Generation 15
2.2 Expected Findings **16**

CHAPTER 3. Results **17**
3.1 Descriptive Statistics **17**
3.2 Measures of Fit **18**
3.3 ANOVA Results **19**
3.3.1 ANOVA Results for RMSE 19
3.3.2 ANOVA Results for Absolute Deviation 25
3.3.3 ANOVA Results for Fit Statistic Estimations 31

CHAPTER 4. Discussion **37**

CHAPTER 5. Summary and Conclusion **40**

REFERENCES **41**

LIST OF TABLES

Table 1	- Descriptive Statistics for Δ^2, item R^2, and Adjusted R^2 Under All Condition Sets	18
Table 2	- Measures of Mean Fit for Δ^2, item R^2, and adjusted R^2 Under All Condition Sets	1919
Table 3	- Three-Way Analysis of Variance of RMSE for Fit Statistic Type, Predictors, and Test Length	20
Table 4	- Three-Way Analysis of Variance of RMSE for Fit Statistic Type, Test Length, and Size of True R^2	21
Table 5	- Three-Way Analysis of Variance of RMSE for Fit Statistic Type, Size of True R^2, and Predictors	22
Table 6	Table 6 - Three-Way Analysis of Variance of Absolute Deviation for Fit Statistic Type, Predictors, and Test Length	26
Table 7	- Three-Way Analysis of Variance of Absolute Deviation for Fit Statistic Type, Test Length, and Size of True R^2	27
Table 8	- Three-Way Analysis of Variance of Absolute Deviation for Fit Statistic Type, Size of True R^2, and Predictors	28
Table 9	- Four-Way Analysis of Variance of Fit Statistic Estimations for Fit Statistic Type, True R^2, Predictors, and Test Length	32

LIST OF FIGURES

Figure 1	- Diagram representation of a single repetition of the process to generate the Δ^2 fit index values using 30 items.	15
Figure 2	- ANOVA Plot for Predictors by Fit Statistic Type for RMSE	23
Figure 3	- ANOVA Plot for Test Length by Fit Statistic Type for RMSE	24
Figure 4	- ANOVA Plot for Size of True R^2 by Fit Statistic Type for RMSE	25
Figure 5	- ANOVA Plot for Predictors by Fit Statistic Type for Absolute Deviation	29
Figure 6	- ANOVA Plot for Test Length by Fit Statistic Type for Absolute Deviation	30
Figure 7	- ANOVA Plot for Size of True R^2 by Fit Statistic Type for Absolute Deviation	31
Figure 8	- ANOVA Plot of Test Length by Fit Statistic Type for Fit Statistic Estimations at True R^2 of 0.36	33
Figure 9	- ANOVA Plot of Test Length by Fit Statistic Type for Fit Statistic Estimations at True R^2 of 0.50	34
Figure 10	- ANOVA Plot of Predictors by Fit Statistic Type for Fit Statistic Estimations at True R^2 of 0.36	35
Figure 11	- ANOVA Plot of Predictors by Fit Statistic Type for Fit Statistic Estimations at True R^2 of 0.50	36

LIST OF SYMBOLS AND ABBREVIATIONS

IRT	Item Response Theory
LLTM	Linear Logistic Test Model
SEM	Structural Equation Modeling
X^2	Chi Squared Statistic
ln	Natural Log
-2lnL	-2 log likelihood value
M_0	General Null Covariance Model Notation from Bentler & Bonett (1980)
M_k	Restricted Covariance Structure Model from Bentler & Bonett (1980)
M_s	Saturated Covariance Structure Model from Bentler & Bonett (1980)
M_t	Target Covariance Structure Model from Bentler & Bonett (1980)
F_0	Null Model Notation Under Bentler and Bonett (1980) General Normed Fit Index Ratio
F_t	Minimum Function Value Notation Correponding to Target Model Under Bentler and Bonett (1980) General Normed Fit Index Ratio
F_k	Minimum Function Value Notation Correponding to Restricted Model Under Bentler and Bonett (1980) General Normed Fit Index Ratio
Δ^2	Delta Squared Index/Delta Squared Statistic
M_N	Null Model Notation in the Δ^2 Statistic
M_S	Saturated Model Notation in the Δ^2 Statistic
M_T	Target Model Notation in the Δ^2 Statistic
β’	IRT True Scores for Items
β	IRT Item estimates
β_T	True Item Difficulty
θ	Person/Subject Ability
η_k	Weight for Predictor k

η_0 Intercept

q_{ik} Predictor k Weight for Item i

SUMMARY

Likelihood ratio chi square tests for nested models are typically used to determine model significance. Multiple correlations of item difficulties estimated with the explanatory predictors are often used to provide further information about model quality. However, the regression approach is not statistically justifiable, since the effective sample size becomes the number of items. Applying explanatory item response theory (IRT) models is advantageous when designing and selecting items. A simulation study was conducted to compare an explanatory item response theory fit statistic, Δ^2 (Embretson, 1997; 2016), to traditionally used fit indices (nested model likelihoods and limited information multiple correlations) for assessing model quality. Test length, item difficulty and the number of predictors were varied and from this, estimations from the Δ^2 as well as item R^2 and adjusted R^2 for comparison were obtained. After computing descriptive statistics and measures of fit (RMSE, bias, and absolute deviation) results support the Δ^2 as a more accurate model for assessing fit over the item R^2 and the adjusted R^2. This simulation study provides needed background for an alternative statistic, Δ^2, for evaluating explanatory IRT models.

CHAPTER 1. INTRODUCTION

Likelihood ratio chi square tests for nested models are typically used to determine model significance. Multiple correlations of item difficulties estimated with the explanatory predictors are often used to provide further information about model quality. However, the regression approach is not statistically justifiable, since the effective sample size becomes the number of items. Applying explanatory item response theory (IRT) models, such as the linear logistic test model (LLTM; Fischer, 1973) is advantageous when designing and selecting items. In addition to providing parameter estimates that identify sources of cognitive complexity in items, explanatory IRT models also provide validity evidence for the response processes aspect. Likelihood ratio chi square tests for nested models are typically used to determine model significance. That is, the explanatory model can be compared to a null model, with equal difficulty for all items, to determine if prediction is significant. Further, as in structural equation modeling, explanatory IRT models can be compared to a saturated model, in which each item receives unique estimates, to determine the adequacy of prediction as illustrated by the chi-square statistic:

$$X^2 = \left(-2lnL_{m_2}\right) - \left(-2lnL_{m_1}\right), \quad (1)$$

where the number of parameters in model 1 (m_1) is greater than the number of parameters in model 2 (m_2). However, both comparisons are usually significant, thus providing little information about model quality. A study published in 2003 by Dimiter Dimitrov and Tenko Raykov discuss this issue when comparing the LLTM fit approach to an SEM

approach and find that the LLTM, when just compared to the fit of the Rasch model, is nearly always significant even when other approaches such as the SEM approach they use actually rejects the LLTM as a good fitting model. For further information on this, Janssen (2016) expands on the technical details of how model fit is fully assessed.

Multiple correlations of item difficulties estimated in the saturated model with the explanatory predictors are often used to provide further information about model quality. However, this approach is not statistically justifiable, since the effective sample size becomes the number of items. In contrast, the Delta statistic (Embretson, 1997; 2016) is based on IRT model log likelihoods. That is, Delta is a ratio comparing the difference in log likelihoods between the null and target models to the difference between the null and saturated models.

The motivation for this fit index results from comparative fit indices and goodness of fit tests, including those analyzed by Bentler and Bonett (1980). The model that Bentler and Bonett analyzed was an incremental fit index, using parameters M_0 to represent a general null model, and M_k, M_t, and M_s to represent hierarchically nested covariance structure models, where M_k is the most restricted model and M_s is the saturated model. These models are typically evaluated relative to each other as chi-square difference tests. As stated, difference tests such as these provide little information in the way of model quality. An incremental fit index (where the index is constrained from zero to one), however, can provide information about practical significance. The more general normed fit index ratio evaluated by Bentler and Bonett is defined as:

$$D_{kt} = (F_k - F_t)/F_0, \quad (2)$$

where F is the -2 log likelihood of an arbitrary model, F_0 is the function evaluated under the null Do model, and F_k and F_t corresponding to the minimum function values for the hierarchically defined step-up models (M_k, M_t) (Bentler & Bonett, 1980). The researchers concluded that finding significant increment in fit would provide evidence that the data are adequate to evaluate the model. While this work in model fit provided motivation for the model to be introduced, a primary difference between how the models are evaluated lies in Bentler and Bonett's, which assumes fit to a correlation matrix of the saturated model as opposed to Embretson's (2016) delta squared (Δ^2) statistic, which directly assumes the saturated model fits raw data.

While the ratio of the Δ^2 statistic is similar in magnitude to a multiple correlation coefficient, its properties have not been compared to this coefficient in a simulation study. Thus, the purpose of this study is to examine how the fit statistic Δ^2 relates to the probability of item difficulties when compared with known true R^2 correlation coefficients, where R is the correlation of IRT true scores (β') to estimates (β), and to estimates of unpredicted sources of item difficulties. The simulation conditions include a) number of predictors, b) test length, c) size of true R^2, and 4) sample size.

1.1 Four Item Response Model Types

Wilson and De Boeck (2004) describe four types of item response theory models; doubly descriptive, person explanatory, item explanatory, and doubly explanatory. These

models illustrate the difference between a descriptive approach and an explanatory approach in item response modeling and will also serve as a basis of laying out the differences between the models chosen for use in the Δ^2 statistic presented later. The models presented are logistic random-intercepts models (Wilson & De Boeck, 2004) and belong to the Rasch tradition, which is also the case for the Δ^2 model.

1.1.1 Doubly Descriptive

A doubly descriptive model is one where each person has a unique effect, unexplained by person properties, and each item has its own unique effect, also unexplained by item properties. A model of this type describes the individual effects without explaining any of the effects. An example of this is the Rasch model, which is described later as one of the models used in the Δ^2 index.

1.1.2 Item Explanatory

An item explanatory model introduces item properties into the Rasch model. The LLTM is an example of an item explanatory model and is one of the models used in the Δ^2 index as well. In the LLTM, item properties are used to explain differences between items and for this reason the contribution of the item on the model is reduced to the contribution of the item properties and the values they have for the item.

1.1.3 Person Explanatory

Similar to the item explanatory model, when person properties are included in the Rasch model, it is referred to as a person explanatory model. The latent regression Rasch model is one example of a person explanatory model and could theoretically be used in the

Δ^2 as the target model in place of the LLTM to examine person differences rather than item differences.

1.1.4 Doubly Explanatory

Finally, as the name implies, a doubly explanatory model includes both item and person properties. Not only are the items and persons being described, but there are properties being included to estimate aspects of the effects that can help provide explanatory information. The latent regression LLTM is an example of this type of model.

1.2 Item Response Theory Models for Analysis

Three IRT models are used as part of the Δ^2 statistic. Two differentiations of the Rasch model is used, one with each item difficulty uniquely estimated and another where the difficulty estimates are denoted as an intercept parameter of the same value. The third model used in the analysis is the LLTM.

1.2.1 Rasch Models

To define our null model (M_N) and saturated model (M_S) in the Delta statistic, the fit index uses Rasch models, where the probability of success for person s on item i is as follows:

$$P(X_{is} = 1|\theta_s, \beta_i) = \frac{e^{1.7(\theta_s - \beta_i)}}{1+e^{1.7(\theta_s - \beta_i)}}, \quad (3)$$

with θ_s denoting the ability of person s and β the difficulty of item i. This particular model represents the saturated model, where every β_i is estimated differently. The null model is nearly identical, apart from β_i, which is denoted as η_0, or some intercept parameter that sets every item difficulty to the same value, as demonstrated in the following model:

$$P(X_{is} = 1|\theta_s, \eta_0) = \frac{e^{1.7(\theta_s - \eta_0)}}{1 + e^{1.7(\theta_s - \eta_0)}}. \quad (4)$$

1.2.2 Linear Logistic Test Model

The linear logistic test model (LLTM) is an extension of the Rasch model, where certain linear constraints are placed on the item parameters. The LLTM is demonstrated by

$$P(X_{is} = 1|\theta_s, \beta_i') = \frac{e^{1.7(\theta_s - \sum \eta_k q_{ik} + \eta_0)}}{1 + e^{1.7(\theta_s - \sum \eta_k q_{ik} + \eta_0)}}, \quad (5)$$

in which θ_s is the ability of person s and β' denotes the predicted item difficulty of $\sum \eta_k q_{ik} + \eta_0$, where η_k is the weight for predictor k, q_{ik} is the predictor k weight for item i, and η_0 is the intercept. The LLTM will act as our target model (M_T) for the delta statistic.

Fischer (1973) demonstrated the usefulness of the LLTM as an instrument in analysis of subject areas in instructional research under the assumption that the subject area comprises tasks or items solved using a combination of a certain number of cognitive operations or rules. Use of the LLTM has continued to increase since Fischer developed the model in 1973. The original purpose of the LLTM was to generate test items with specified item difficulties; in fact, Fischer and Formann (1982) reference several early

studies (including Fischer's 1973 study) in which the LLTM is proven to be useful for item analysis, but that the model only attains good fit if the item material was constructed carefully and with the model in mind. They also reference a study by Formann (1973), which sought to create a new nonverbal intelligence test using items designed from predetermined construction rules. Formann's (1973) study revealed that the difficulty of new items can be predicted or tasks with prespecified difficulties can be constructed using the parameter estimates of the elementary operations, since a set of rules governing the items were in place to allow for structural or superficial differences while still allowing for difficulties to be predicted.

There have also been several studies providing examples of other applications of the LLTM in psychometric research in addition to its capabilities in item generation. Whitely and Schneider (1981), for example, utilized the LLTM in assessing item bias of gender by using geometric analogy items from the Cognitive Abilities Test (Thorndike & Hagen, 1974). Different aspects of testing conditions such as position effects, learning and fatigue effects (specific types of position effects), speeded item presentation effects, and item response format effects can also be carried out using LLTM. An adaptive test called the Adaptive Intelligence Diagnosticum, version 2.1 (Kubinger & Wurst, 2000) was, in fact, used to demonstrate the application of the LLTM to all aforementioned testing conditions (Kubinger, 2009). Another application of note is incorporating cognitive complexity into latent trait models (including the LLTM). Embretson (1994), using items from the Spatial Learning Ability Test (Embretson & Waxman, 1989), attempted to bridge cognitive psychology with psychometric testing by conceptualizing traits in cognitive models as latent factors and fitting these cognitive models to the LLTM to assess ability.

Perhaps most relevant to the current study is a study from Susan Embretson and Robert Daniel published in 2008 where they apply both LLTM and regression modeling to mathematical items from the quantitative section of the GRE. Results from Embretson & Daniel (2008) support the LLTM as a more consistent and powerful estimator of the impact of variable complexity in their cognitive model than the regression modeling approach.

While these studies do provide many potential applications to LLTM, many of these studies also warn that because of the method of obtaining a χ^2 statistic, significant values are common. In addition, none of these applications fully assess model quality, which reiterates the need for the current study.

1.2.3 Fit Statistic Models

Three fit statistic models will be compared in this thesis: R^2 statistic, adjusted R^2 statistic, and the Δ^2 statistic. The following section details the three models.

1.2.3.1 R Squared Statistic

The R^2 statistic, or coefficient of determination, represents the proportion of variance in one variable (e.g., dependent variable y) as explained by some other variables (e.g., independent variables x_1 to x_n). The way this is calculated in a regression context is by taking the amount of variability that is explained by the independent variables and creating a proportion over the total amount of variation, illustrated by equation 6. For the purposes of clarity between similarly named terms, this statistic will be referred to as the item R^2 through the rest of the thesis.

$$R^2 = \frac{Var(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_n X_n)}{Var(Y)} \quad (6)$$

1.2.3.2 Adjusted R Squared Statistic

The adjusted R^2 is a modified version of the R^2 statistic that is adjusted for the number of predictors in the model. In other words, the proportion of variance explained only increases when the predictor added to the model is statistically significant. The R^2 statistic is adjusted according to the following,

$$R^2_{Adj} = \frac{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - y_i)^2} \quad (7)$$

Where n is the sample size and p is the number of predictors in the model. This equation is advantageous over R^2 with a larger number of predictors since it only increases when the predictors are significant. This also means the proportion of variance for the adjusted R^2 relative to R^2 testing the same model will always be equal to or less than R^2.

1.2.3.3 Delta Squared Statistic

The Δ^2 fit statistic is a full information comparison of likelihoods of three models, with each item containing scores on the predictors, as demonstrated by the equation:

$$\Delta^2 = \frac{(-2\ln L_{M_N}) - (-2\ln L_{M_T})}{(-2\ln L_{M_N}) - (-2\ln L_{M_S})}, \quad (8)$$

where $\ln L_{M_N}$, $\ln L_{M_T}$ and $\ln L_{M_S}$ are the -2-log likelihoods for the null, target and saturated models, respectively. The saturated model in our statistic is represented by the Rasch model, the target model is the explanatory model of item predictors, such as provided with the LLTM, and the null model is represented by a Rasch model, in which the item difficulties are equal for all items and essentially act as an intercept parameter.

While the nested likelihood model produces an asymptotic chi-squared distribution, the Δ^2 statistic is a ratio of two chi-squared models, giving a value that can be interpreted similar to a squared multiple correlation coefficient. Considering Bentler and Bonett's (1980) nested framework, the Δ^2 statistic can be nested in a similar manner for comparison. The saturated model contains the most information of the three models used, since it is calculated using a Rasch model where each difficulty has a unique estimate. Nested within the saturated model would be the target model, calculated using the LLTM. The target model contains less information than the Rasch model since it has a set of difficulty parameters that are estimated rather than using a different difficulty estimate for each item. Finally, the null model, nested within the target model, is calculated using a Rasch model where the difficulty parameters are defined as some intercept value, meaning the model provides the least information.

CHAPTER 2. METHODS

Data are simulated under various conditions to evaluate the Δ^2 fit index. The conditions and simulation process for the Δ^2 fit index is derived from normal metric IRT models representing null, target, and saturated models using specialized macros developed for IBM SPSS statistical software, while the log likelihood values to calculate the model were estimated using SAS. The estimates obtained from the three models are used to compute the Δ^2 statistic, which is then compared to the true item bank R^2, observed limited information (item) R^2, and adjusted R^2 to assess fit. Parameter estimates are evaluated with respect to their reliability in IRT standard deviation of the estimates, bias, and root mean square error (RMSE) approximations. ANOVA tables and plots were obtained to verify the accuracy of the obtained results.

2.1 Data Simulation

The generating and calibration codes were developed using SPSS. First, an SPSS macro file was written to generate item responses by comparing the probability of item success from true difficulty (β_T) and subject's ability level (θ) to a random selection from a uniform distribution $U(0\text{-}1)$. Second, the item selection and generation code were developed to select items for each replication in each condition.

To derive the target model estimates, a separate set of code was developed that derives true item difficulties (β_T) from a single predictor model and can be expanded to include multiple predictors (namely 5 and 8 predictors) for the assessment (see equation 11). The predictors are generated under random normal distribution, $N(0, 1)$. The error term in these

models is always represented as a random normal variable. Using these estimates, binary person responses were developed for each of the items and placed in long vector format along with q values from our LLTM model calculations and a set of either 20 or 30 dummy codes, depending on our condition, for computing log likelihoods and beta estimates in SAS that are used in the analysis.

2.1.1 Data Conditions

Data are generated for the normal metric variant of the Rasch model with theta as a normal random distribution and true item difficulty based on uncorrelated predictors plus prediction error. Responses are simulated for 300 subjects per condition, where the conditions vary in the number of predictors (5 or 8 predictors), the test length consisting of 20 or 30 items, and the size of the true R^2 (0.36 or 0.5). The predictors are estimated as random normal variables. When selecting items for the test length condition, the items are generated as individual item banks of 300 items and, for each repetition, a subset of items are selected from the dataset corresponding to the repetition.

When choosing the number of predictors, the motivation was to choose two conditions that reasonably demonstrated differences between the conditions. Additionally, the difference in the number of predictors is small (3 predictor difference between the two conditions) which will further support significant results. For test length conditions, 30 items were chosen because that is the average length of many tests, so it is expected that using a test length of 30 items is a good indicator of model performance. In the case of 20 items, the trade-off between accuracy and test times have an important question for researchers and test developers are constantly considering how to develop items/tests to

allow for the most accurate test that can be developed with the least amount of questions. The reason for assessing the model with the 20 item condition is to assess the capability of the model with shorter tests.

The three simulation conditions yield a total of 8 conditions (i.e., 2 Predictors x 2 Test Length x 2 True R^2). Each condition is replicated 100 times, generating a total of 800 datasets. Items for each replication are selected from the test banks of 300 items, with each replication randomly selecting either 20 or 30 items, for each predictor condition.

2.1.2 Dependent Measures

The dependent measures include RMSE, bias, and the fit statistic estimations. The fit statistic estimations are the calculated values from Δ^2, item R^2, and the adjusted R^2 so the purpose of including this as a dependent variable was to look at the difference of these estimates among each predictor or among different interactions of predictors.

The RMSE value is a measure of consistency and examines the amount of error there is around the target value. More specifically, the RMSE gives an indication of whether the set of estimations consistently estimates similar values, hopefully near the true R^2 condition. RMSE is defined by the equation,

$$RMSE = \sqrt{\left(\frac{F - R_T^2}{n}\right)^2}, \tag{9}$$

where F denotes the estimated fit statistic being examined and R_T^2 denotes the true R^2 value corresponding to the respective condition.

Bias measures accuracy of an estimate, and optimal bias values (those that are close to zero) in this case are those where the estimates are all close to its respective true R^2 values. The equation for bias is defined as,

$$Bias = \frac{F - R_T^2}{n}. \tag{10}$$

Since bias produces both positive and negative results around the ideal point of zero, this equation was modified by first taking the absolute value of the difference before taking the mean, thus creating a measure of absolute deviation. This absolute deviation statistic was used to visually compare the fit statistics, while the bias values were used in the analyses.

2.1.3 Generating True Item Difficulties

Code was developed for true item difficulties (β_T) from a specified target model as follows:

$$\beta_T = \eta_1 q_1 + \eta_2 q_2 + \cdots + \eta_n q_n + \eta_{error}, \tag{11}$$

where η_1 to η_n sum to β, q_1 to q_n are standard normal predictors, and η_{error} is the square root of one minus the sum of the squared η_n weights.

2.1.4 Item Response Generation

The first step for generation in a condition is randomly selecting items, as described above. Individual datasets of 1,000 subjects for each repetition are generated for a total of 100,000 subjects, with theta ~*N* (0,1), and item responses are generated for the items by computing expected probabilities from the true item difficulties and comparing to a selection from a random uniform distribution (range 0 to 1). Both item sampling and subject sampling are done without replacement. Figure 1 visually describes a single repetition of the generation process with 30 items.

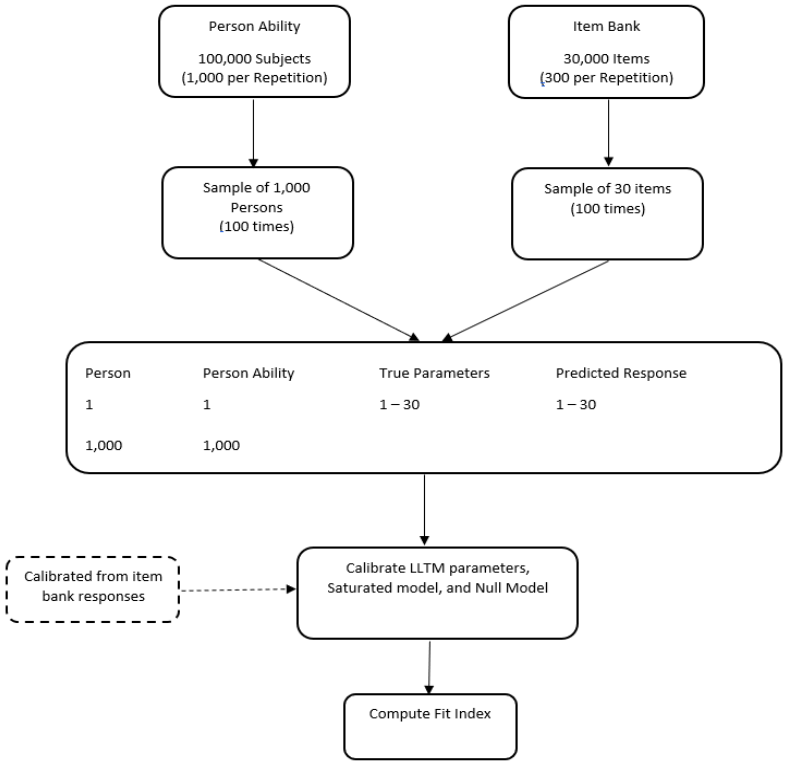

Figure 1 - Diagram representation of a single repetition of the process to generate the Δ^2 fit index values using 30 items.

2.2 Expected Findings

The results will be comparing the Δ^2 statistic to R^2 distributions to examine which gives better estimates. Since the Δ^2 statistic is based on the full information from the item response data, while the observed R^2 is based only on the item parameters, it is expected that Δ^2 is a better estimator for item predictability. The R^2 estimates are also expected to be more vulnerable to sampling from the item bank of 300 to just 30 items. In addition, with varying conditions, Δ^2 and the true R^2 are likely to be closer with more items, fewer predictors, and larger R^2 values (e.g., five predictors, 30 items, and an $R^2 = 0.5$).

CHAPTER 3. RESULTS

The NLMIXED procedure in SAS was used to derive log likelihood values and beta estimates. The beta estimates were used to compare to the true betas calculated from the saturated model by regressing the estimated item difficulties obtained from our saturated model on the q values calculated for our true item difficulty model. The log likelihoods were used to compute the Δ^2 statistic.

3.1 Descriptive Statistics

Descriptive statistics are displayed in Table 1 for each model type. The columns for mean display the values that attempt to estimate the true R^2 relative to its respective condition set. The means show that for the Δ^2 statistic, the estimates were closest to the true R^2 value only in condition sets one and five, while the estimates were closest to true R^2 in the remaining condition sets for the adjusted R^2. However, the columns displaying the standard deviation of the estimates reveal that the adjusted R^2 has the highest standard deviation value for every condition set. The Δ^2 and item R^2 standard deviation estimates are lowest in four condition sets each.

Table 1 - Descriptive Statistics for Δ^2, item R^2, and Adjusted R^2 Under All Condition Sets

	Mean			Standard Deviation		
Condition Set	*Delta Squared*	*Item R^2*	*Adjusted R^2*	*Delta Squared*	*Item R^2*	*Adjusted R^2*
20 items; 5 parms; True $R^2 = 0.5$	0.517	0.590	0.444	0.150	0.140	0.190
20 items; 5 parms; True $R^2 = 0.36$	0.437	0.516	0.343	0.149	0.155	0.210
20 items; 8 parms; True $R^2 = 0.5$	0.635	0.695	0.473	0.128	0.125	0.215
20 items; 8 parms; True $R^2 = 0.36$	0.568	0.634	0.368	0.138	0.127	0.220
30 items; 5 parms; True $R^2 = 0.5$	0.503	0.576	0.488	0.128	0.130	0.157
30 items; 5 parms; True $R^2 = 0.36$	0.387	0.460	0.348	0.120	0.149	0.149
30 items; 8 parms; True $R^2 = 0.5$	0.549	0.625	0.482	0.122	0.112	0.155
30 items; 8 parms; True $R^2 = 0.36$	0.441	0.518	0.335	0.1308	0.1314	0.182

3.2 Measures of Fit

Table 2 contains the values for RMSE, bias, and absolute deviation. The RMSE values are lowest in all conditions for Δ^2 with the exception of condition 4 (20 items; 8 parms; True $R^2 = 0.36$). On the other hand, the adjusted R^2 contains the most conditions where bias is closest to zero. However, looking at the absolute deviation values indicates that the bias may be misleading, since the absolute deviation indicates that most of the Δ^2 values end up closest zero.

Table 2 - Measures of Mean Fit for Δ^2, item R^2, and adjusted R^2 Under All Condition Sets

	RMSE			Bias			Absolute Deviation		
Condition Set	*Delta Squared*	*Item R^2*	*Adjusted R^2*	*Delta Squared*	*Item R^2*	*Adjusted R^2*	*Delta Squared*	*Item R^2*	*Adjusted R^2*
20 items; 5 parms; True R^2 = 0.5	0.150	0.166	0.197	0.017	0.090	-0.056	0.120	0.140	0.149
20 items; 5 parms; True R^2 = 0.36	0.167	0.219	0.209	0.077	0.156	-0.170	0.136	0.183	0.173
20 items; 8 parms; True R^2 = 0.5	0.186	0.213	0.216	0.135	0.195	-0.027	0.156	0.202	0.177
20 items; 8 parms; True R^2 = 0.36	0.249	0.301	0.219	0.208	0.274	0.008	0.218	0.275	0.183
30 items; 5 parms; True R^2 = 0.5	0.127	0.150	0.157	0.003	0.076	-0.012	0.103	0.125	0.123
30 items; 5 parms; True R^2 = 0.36	0.123	0.158	0.148	0.027	0.100	-0.012	0.098	0.131	0.123
30 items; 8 parms; True R^2 = 0.5	0.131	0.167	0.155	0.049	0.125	-0.018	0.107	0.144	0.126
30 items; 8 parms; True R^2 = 0.36	0.153	0.205	0.182	0.081	0.158	-0.025	0.126	0.179	0.144

3.3 ANOVA Results

ANOVA results were obtained using RMSE, absolute deviation, and fit statistic estimations as dependent variables. When analyzing RMSE and absolute deviation, only three-way ANOVA results could be obtained since there was too little variability; because of this, three different ANOVA combinations were run to ensure all combinations of predictors were analyzed. However, a full four-way ANOVA was able to be performed on the fit statistic estimations, so only one table was obtained for this result. Significance tables and plots of marginal means were obtained for all three dependent variables.

3.3.1 ANOVA Results for RMSE

The first combination of predictors in Table 3 assess variance of RMSE for fit statistic type, predictors, and test length. According to the results, all predictor conditions are significant. Fit Statistic Type x Test Length was significant at $p < 0.05$ and all other conditions were significant at $p < 0.001$.

Table 3 - Three-Way Analysis of Variance of RMSE for Fit Statistic Type, Predictors, and Test Length

Source	Type III SS	df	Mean Square	F	significance
Total Model	3.188*	11	0.290	778.517	<0.001
Intercept	78.771	1	78.771	211568.760	<0.001
Fit Statistic Type	0.558	2	0.279	749.098	<0.001
Predictors	0.687	1	0.687	1844.695	<0.001
Test Length	1.685	1	1.685	4526.752	<0.001
Fit Statistic Type x Predictors	0.114	2	0.587	153.240	<0.001
Fit Statistic Type x Test Length	0.003	2	0.002	4.264	<0.05
Predictors x Test Length	0.089	1	0.089	238.549	<0.001
Fit Statistic Type x Predictors x Test Length	0.052	2	0.026	70.246	<0.001
Error	0.889	2388	0		
Total	82.849	2400			
Corrected Total	4.078	2399			

* R Squared = 0.782; Adjusted R Squared = 0.781

Table 4 shows the results assessing variance of RMSE for fit statistic type, test length, and size of the true R^2. These results are also significant for all predictors. Fit Statistic Type x Test Length is significant at $p < 0.05$ and all other conditions are significant at $p < 0.001$.

Table 4 - Three-Way Analysis of Variance of RMSE for Fit Statistic Type, Test Length, and Size of True R^2

Source	Type III SS	df	Mean Square	F	significance
Total Model	2.978*	11	0.271	588.071	<0.001
Intercept	78.771	1	78.771	171098.585	<0.001
Fit Statistic Type	0.558	2	0.279	605.806	<0.001
Test Length	1.685	1	1.685	3660.847	<0.001
True R^2	0.421	1	0.421	915.212	<0.001
Fit Statistic Type x Test Length	0.003	2	0.002	3.448	<0.05
Fit Statistic Type x True R^2	0.149	2	0.075	162.283	<0.001
Test Length x True R^2	0.099	1	0.099	214.639	<0.001
Fit Statistic Type x Test Length x True R^2	0.062	2	0.031	67.507	<0.001
Error	1.099	2388	0		
Total	82.849	2400			
Corrected Total	4.078	2399			

* R Squared = 0.730; Adjusted R Squared = 0.729

Results for Table 5 that assess variance of RMSE using the predictors fit statistic type, size of true R^2, and number of predictors show a similar trend as the previous two tables. In this case, all results are significant at $p < 0.001$.

Table 5 - Three-Way Analysis of Variance of RMSE for Fit Statistic Type, Size of True R^2, and Predictors

Source	Type III SS	df	Mean Square	F	significance
Total Model	2.056*	11	0.187	220.843	<0.001
Intercept	78.771	1	78.771	93061.784	<0.001
Fit Statistic Type	0.558	2	0.279	329.502	<0.001
True R^2	0.421	1	0.421	497.790	<0.001
Predictors	0.687	1	0.687	811.418	<0.001
Fit Statistic Type x True R^2	0.149	2	0.075	88.267	<0.001
Fit Statistic Type x Predictors	0.114	2	0.057	67.405	<0.001
True R^2 x Predictors	0.112	1	0.112	132.398	<0.001
Fit Statistic Type x True R^2 x Predictors	0.015	2	0.007	8.659	<0.001
Error	2.021	2388	0.001		
Total	82.849	2400			
Corrected Total	4.078	2399			

* R Squared = 0.504; Adjusted R Squared = 0.502

As indicated in Table 2, the plots for RMSE indicate that the Δ^2 statistic contains the lowest values overall. The item R^2 and adjusted R^2 are close in RMSE for the five predictor condition, with the item R^2 slightly lower. However, for the eight predictor condition, the RMSE for item R^2 is much higher.

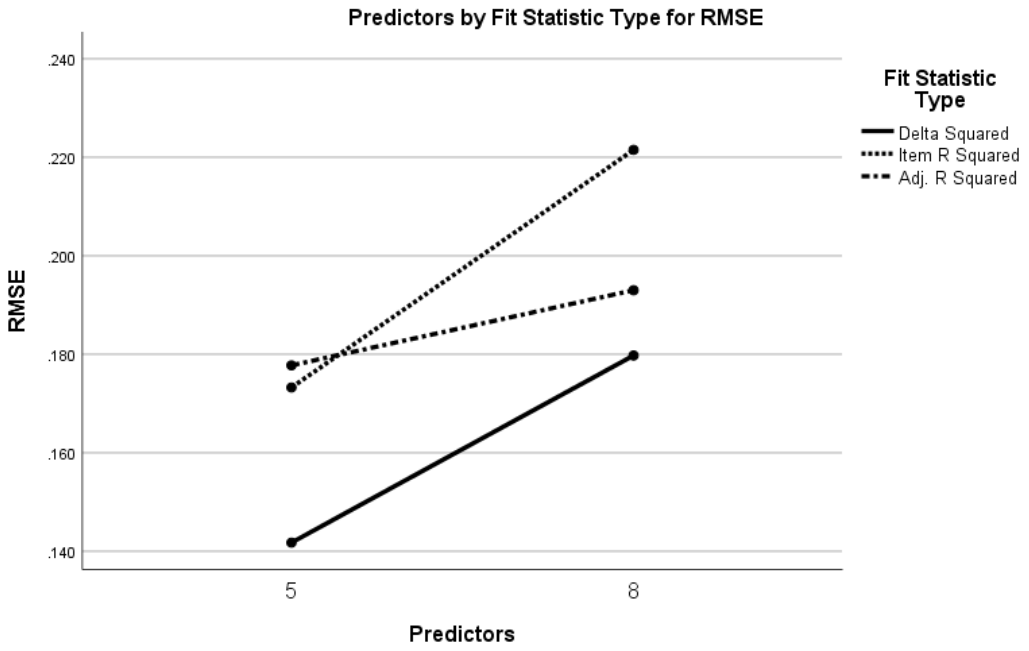

Figure 2 - ANOVA Plot for Predictors by Fit Statistic Type for RMSE

The three fit statistics follow the same trend for test length, all having higher RMSE for the test length of 20 items and much lower RMSE for 30 items. Again, the Δ^2 is lower for both conditions and the item R^2 is highest for both conditions.

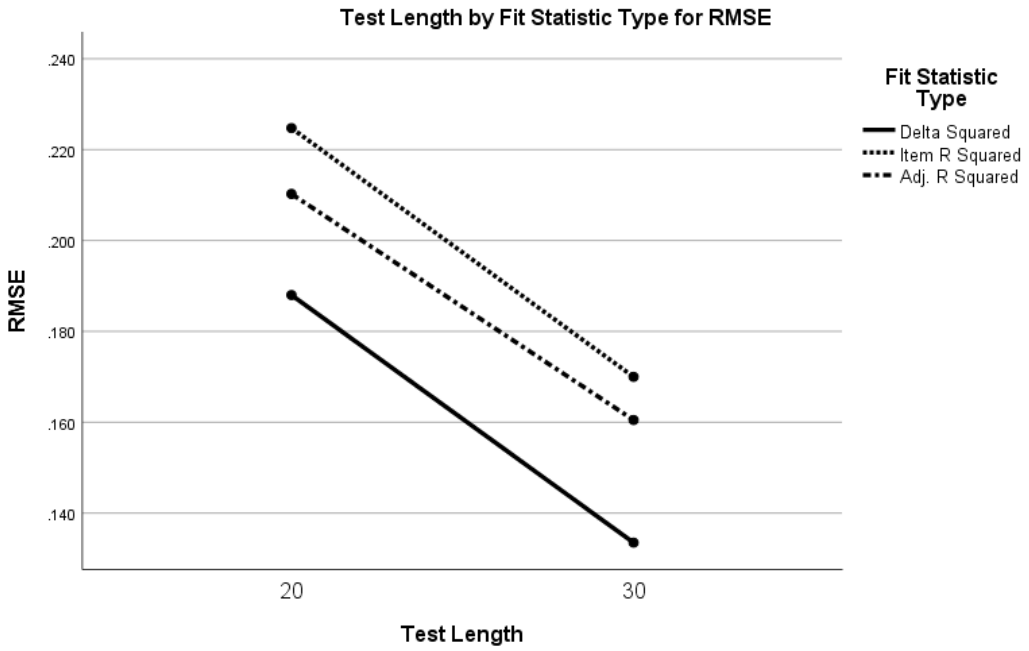

Figure 3 - ANOVA Plot for Test Length by Fit Statistic Type for RMSE

The RMSE is again higher for all fit statistic types for the size of the true R^2 at 0.36 and gets lower for all fit statistics when true R^2 is 0.5. Like the previous plots, the Δ^2 statistic is lowest for both conditions. When the true R^2 is 0.36, the RMSE for adjusted R^2 is much lower than the item R^2. However, the adjusted R^2 does not decrease much for the other condition so when the true R^2 is at 0.5, the item R^2 actually has a slightly lower RMSE than the adjusted R^2 as depicted in Figure 4.

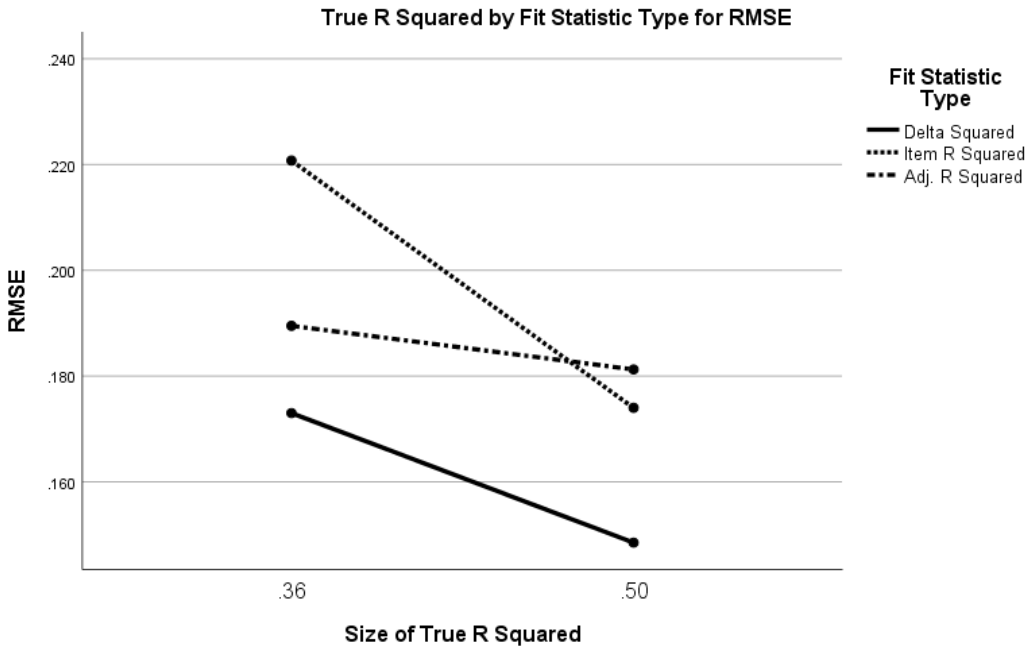

Figure 4 - ANOVA Plot for Size of True R^2 by Fit Statistic Type for RMSE

3.3.2 ANOVA Results for Absolute Deviation

Table 6 shows the ANOVA results of absolute deviation for fit statistic type, predictors, and test length. As indicated in the table, all predictors were significant at $p < 0.001$.

Table 6 - Three-Way Analysis of Variance of Absolute Deviation for Fit Statistic Type, Predictors, and Test Length

Source	Type III SS	df	Mean Square	F	significance
Total Model	3.189*	11	0.290	998.999	< 0.001
Intercept	55.237	1	55.237	190326.767	< 0.001
Fit Statistic Type	0.625	2	0.312	1076.346	< 0.001
Predictors	0.781	1	0.781	2691.747	< 0.001
Test Length	1.416	1	1.416	4879.728	< 0.001
Fit Statistic Type x Predictors	0.159	2	0.079	273.254	< 0.001
Fit Statistic Type x Test Length	0.019	2	0.009	32.662	< 0.001
Predictors x Test Length	0.146	1	0.146	502.044	< 0.001
Fit Statistic Type x Predictors x Test Length	0.044	2	0.022	75.474	< 0.001
Error	0.693	2388	0.000		
Total	59.119	2400			
Corrected Total	3.882	2399			

* R Squared = 0.821; Adjusted R Squared = 0.821

Results assessing variance of absolute deviation for fit statistic type, test length, and size of true R^2 are contained in Table 7. Again, all of the predictors and interactions are significant at $P < 0.001$.

Table 7 - Three-Way Analysis of Variance of Absolute Deviation for Fit Statistic Type, Test Length, and Size of True R^2

Source	Type III SS	df	Mean Square	F	significance
Corrected Model	2.626*	11	0.239	453.750	< 0.001
Intercept	55.237	1	55.237	104991.416	< 0.001
Fit Statistic Type	0.625	2	0.312	593.753	< 0.001
Test Length	1.416	1	1.416	2691.842	< 0.001
True R^2	0.368	1	0.368	698.595	< 0.001
Fit Statistic Type x Test Length	0.019	2	0.009	18.017	< 0.001
Fit Statistic Type x True R^2	0.075	2	0.038	71.444	< 0.001
Test Length x True R^2	0.095	1	0.095	180.579	< 0.001
Fit Statistic Type x Test Length x True R^2	0.028	2	0.014	26.903	< 0.001
Error	1.256	2388	0.001		
Total	59.119	2400			
Corrected Total	3.882	2399			

* R Squared = 0.676; Adjusted R Squared = 0.675

ANOVA results of absolute deviation for fit statistic type, size of true R^2, and predictors are in Table 8. Again, results indicate all interactions and marginal means are significant at $p < 0.001$.

Table 8 - Three-Way Analysis of Variance of Absolute Deviation for Fit Statistic Type, Size of True R^2, and Predictors

Source	Type III SS	df	Mean Square	F	significance
Corrected Model	2.112*	11	0.192	259.006	<0.001
Intercept	55.237	1	55.237	74512.620	<0.001
Fit Statistic Type	0.625	2	0.312	421.388	<0.001
True R^2	0.368	1	0.368	495.794	<0.001
Predictors	0.781	1	0.781	1053.815	<0.001
Fit Statistic Type x True R^2	0.075	2	0.038	50.704	<0.001
Fit Statistic Type x Predictors	0.159	2	0.079	106.978	<0.001
True R^2 x Predictors	0.069	1	0.069	93.534	< 0.001
Fit Statistic Type x True R^2 x Predictors	0.035	2	0.018	23.893	< 0.001
Error	1.770	2388	0.001		
Total	59.119	2400			
Corrected Total	3.882	2399			

* R Squared = 0.544; Adjusted R Squared = 0.542

The plot for predictor by fit statistic type for absolute deviation in Figure 5 shows Δ^2 as much lower than both item R^2 and adjusted R^2 for the 5 predictor condition. Although the slope for Δ^2 is steeper than that of the adjusted R^2, its absolute deviation value still remains below adjusted R^2 for the 8 predictor condition. Item R^2 absolute deviation was nearly the same as the adjusted R^2 value for the 5 predictor condition, but steeply increased

when estimating 8 predictors, to nearly 1.5 times higher absolute deviation value than the other two statistics.

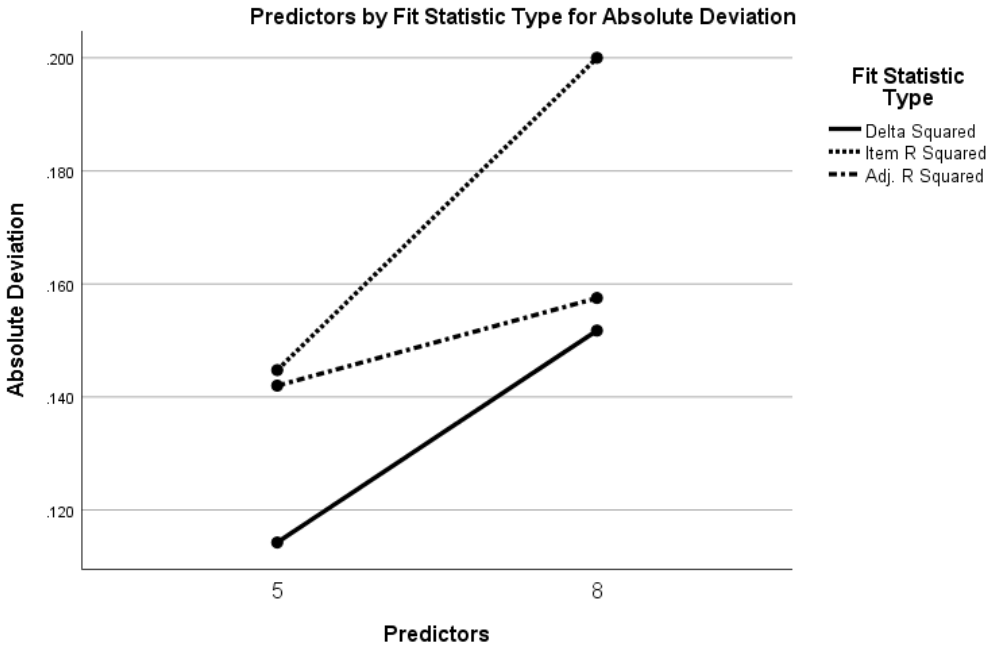

Figure 5 - ANOVA Plot for Predictors by Fit Statistic Type for Absolute Deviation

Figure 6 depicts the test length conditions by fit statistic types for absolute deviation. All statistics have a higher absolute deviation for the 20 item condition and decrease for 30 items. Additionally, all have slopes that are decreasing at approximately the same rate. Δ^2 does have the lowest absolute deviation estimations for both the 20 and 30 item conditions, and examining the plot closely, it can be seen that the slope for Δ^2 appears to be slightly steeper than the slope for adjusted R^2.

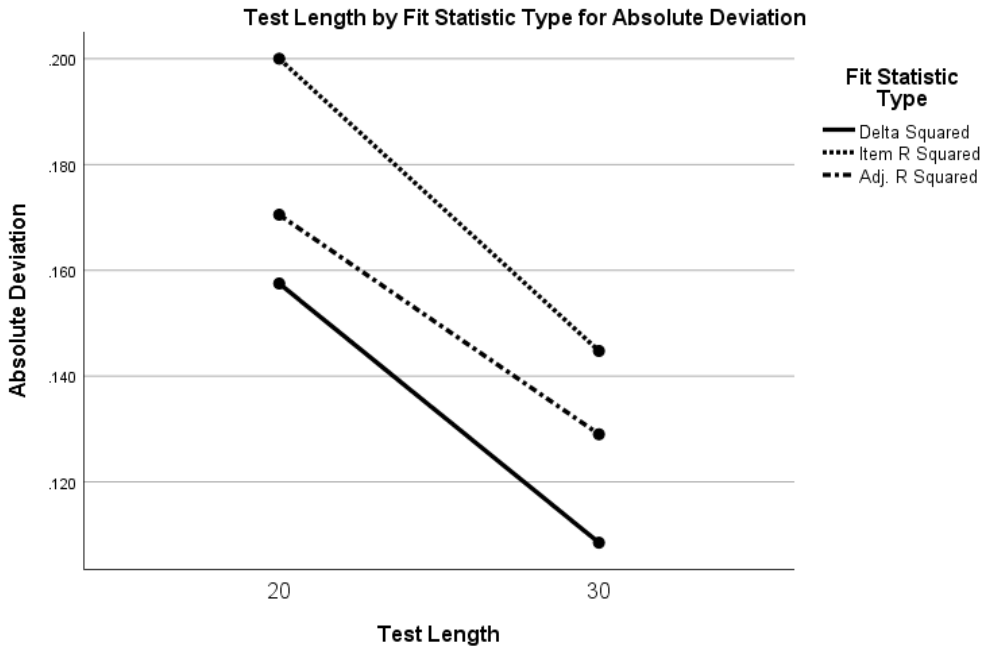

Figure 6 - ANOVA Plot for Test Length by Fit Statistic Type for Absolute Deviation

Plot for true R^2 by fit statistic type for absolute deviation in Figure 7 shows Δ^2 with the lowest absolute deviation estimates for both true R^2 sizes as well as a much steeper slope than the adjusted R^2. Item R^2 has a steep slope than the adjusted R^2 as well, but much higher absolute deviation values for both conditions, so its slope values are not enough to estimate lower that adjusted R^2, although item R^2 does approximate close to adjusted R^2 absolute deviation values for the condition where the size of true $R^2 = 0.5$.

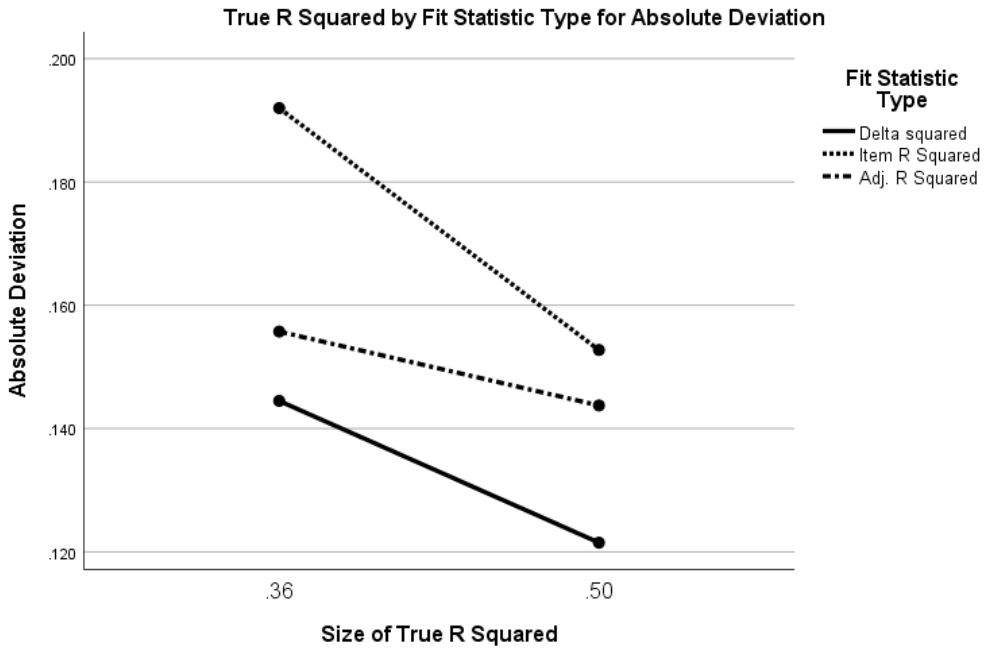

Figure 7 - ANOVA Plot for Size of True R^2 by Fit Statistic Type for Absolute Deviation

3.3.3 ANOVA Results for Fit Statistic Estimations

Table 9 shows the four-way ANOVA of fit statistic estimations for our variables. All individual variables were significant and the only two-way interaction that was non-significant is the true R^2 x predictors interaction. Additionally, all three-way interactions were statistically non-significant as well as the four-way interaction.

Table 9 - Four-Way Analysis of Variance of Fit Statistic Estimations for Fit Statistic Type, True R^2, Predictors, and Test Length

Source	Type III SS	df	Mean Square	F	significance
Corrected Model	23.017*	23	1.001	42.997	< 0.001
Intercept	593.590	1	593.590	25503.447	< 0.001
Fit Statistic Type	11.180	2	5.590	240.176	< 0.001
True R^2	6.235	1	6.235	267.899	< 0.001
Predictors	2.125	1	2.125	91.288	< 0.001
Test Length	1.078	1	1.078	46.319	< 0.001
Fit Statistic Type x True R^2	0.141	2	0.070	3.020	< 0.05
Fit Statistic Type x Predictors	0.771	2	0.385	16.554	< 0.001
Fit Statistic Type x Test Length	0.706	2	0.353	15.176	< 0.001
True R^2 x Predictors	0.004	1	0.004	0.191	0.662
True R^2 x Test Length	0.252	1	0.252	10.811	< 0.01
Predictors x Test Length	0.479	1	0.479	20.563	< 0.001
Fit Statistic Type x True R^2 x Predictors	0.009	2	0.004	0.190	0.827
Fit Statistic Type x True R^2 x Test Length	0.001	2	0	0.014	0.986
Fit Statistic Type x Predictors x Test Length	0.037	2	0.018	0.789	0.454
True R^2 x Predictors x Test Length	0.001	1	0.001	0.027	0.869
Fit Statistic Type x True R^2 x Predictors x Test Length	1.466×10^{-5}	2	7.330×10^{-6}	0	1
Error	55.301	2376	0.023		
Total	671.909	2400			
Corrected Total	78.319	2399			

* R Squared = 0.294; Adjusted R Squared = 0.287

Figure 8 depicts test length for fit statistic type at true R^2 of 0.36. The plot indicates a negative trend as test length increases. Adjusted R^2 is approximately close to 0.36 for both test length conditions, but still trends down as test length increases. Δ^2 is high at an average of around 0.5 for test length of 20 items but corrects heavily for the 30 item condition to an average of around 0.41.

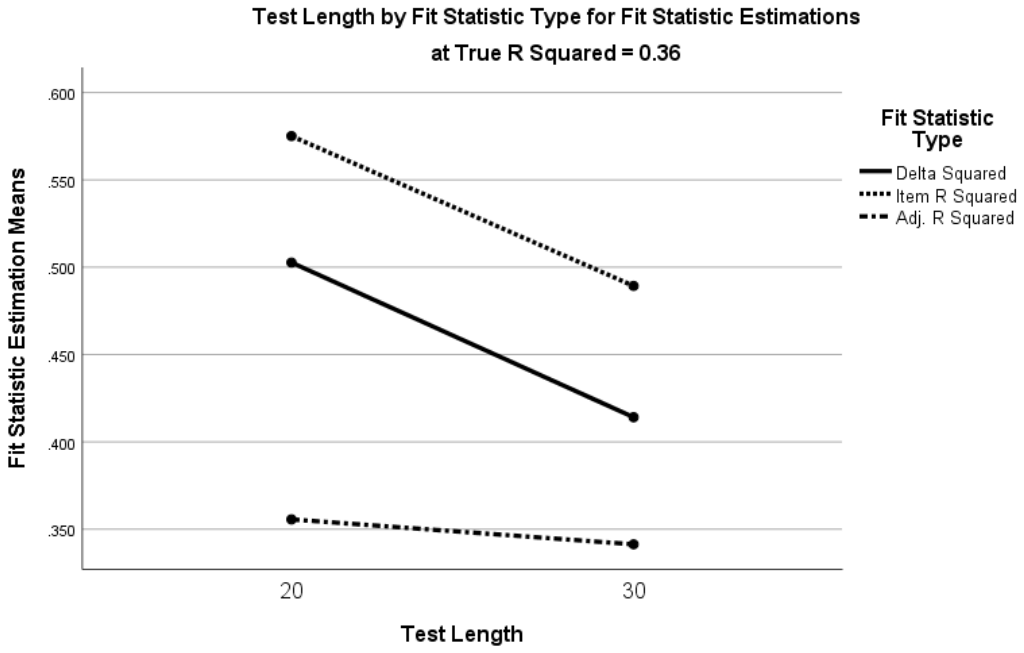

Figure 8 - ANOVA Plot of Test Length by Fit Statistic Type for Fit Statistic Estimations at True R^2 of 0.36

Figure 9 depicts test length for the fit statistics at true R^2 of 0.5. For the 20 item condition, adjusted R^2 has an average of 0.46 and corrects upward to an average of approximately 0.48 for the 30 item condition. Δ^2 has an average of around 0.57, correcting to approximately 0.525 for 20 and 30 items, respectively. Item R^2 average remained at 0.6 and higher.

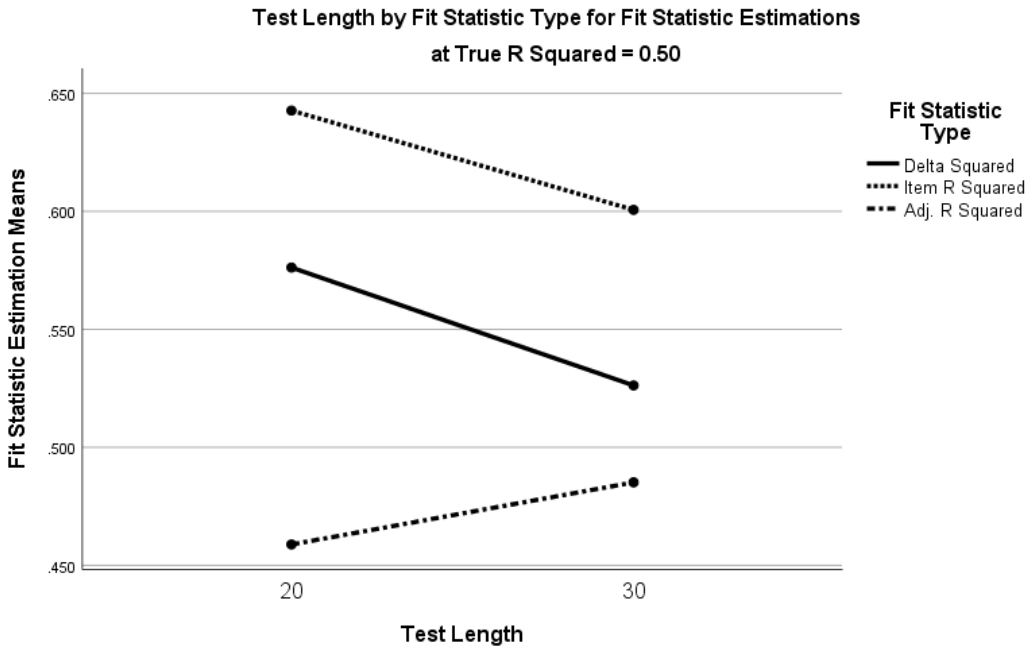

Figure 9 - ANOVA Plot of Test Length by Fit Statistic Type for Fit Statistic Estimations at True R^2 of 0.50

Figure 10 shows the plot for predictors by fit statistic type for the estimations at the true R^2 of 0.36. Adjusted R^2 is close to the average in both conditions, hovering around 0.35, while estimation means for both Δ^2 and item R^2 increase as number of predictors increase. The average for item R^2 remains highest in both conditions.

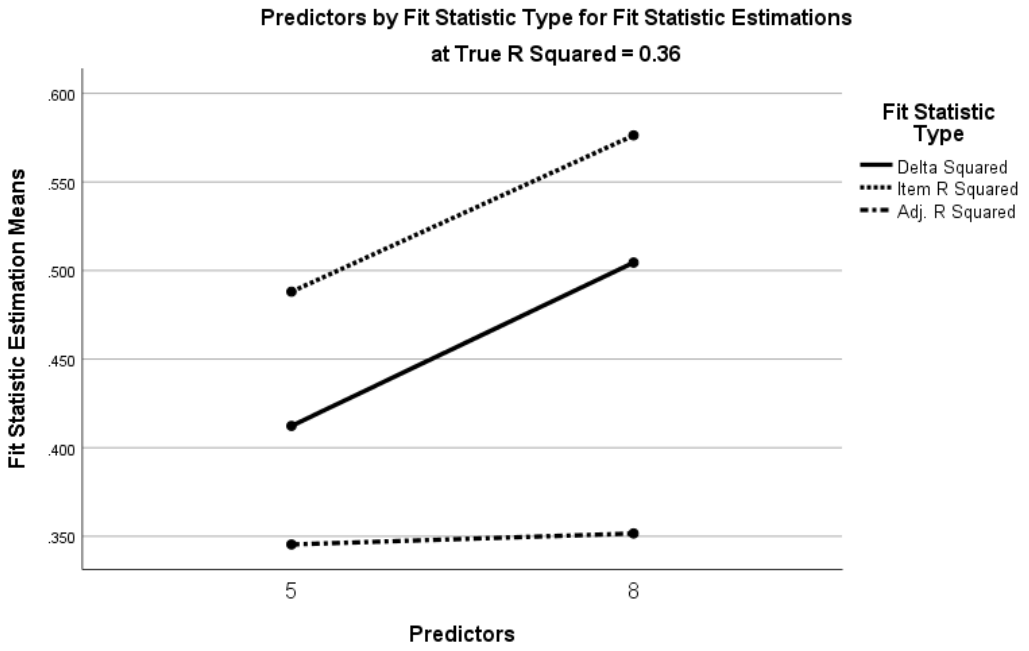

Figure 10 - ANOVA Plot of Predictors by Fit Statistic Type for Fit Statistic Estimations at True R^2 of 0.36

Plot for predictors by fit statistic type for the estimations at true R^2 of 0.5 is depicted in Figure 11. For all three fit statistics, as number of predictors increase, the estimation means increase. The adjusted R^2 remains closest to the 0.5 average, with less of a slope than the other two statistics. Both Δ^2 and item R^2 increase sharply as number of predictors increase. However, the Δ^2 estimated at approximately 0.5 and increases to approximately 0.6 while the item R^2 average is higher even at the lower predictor condition.

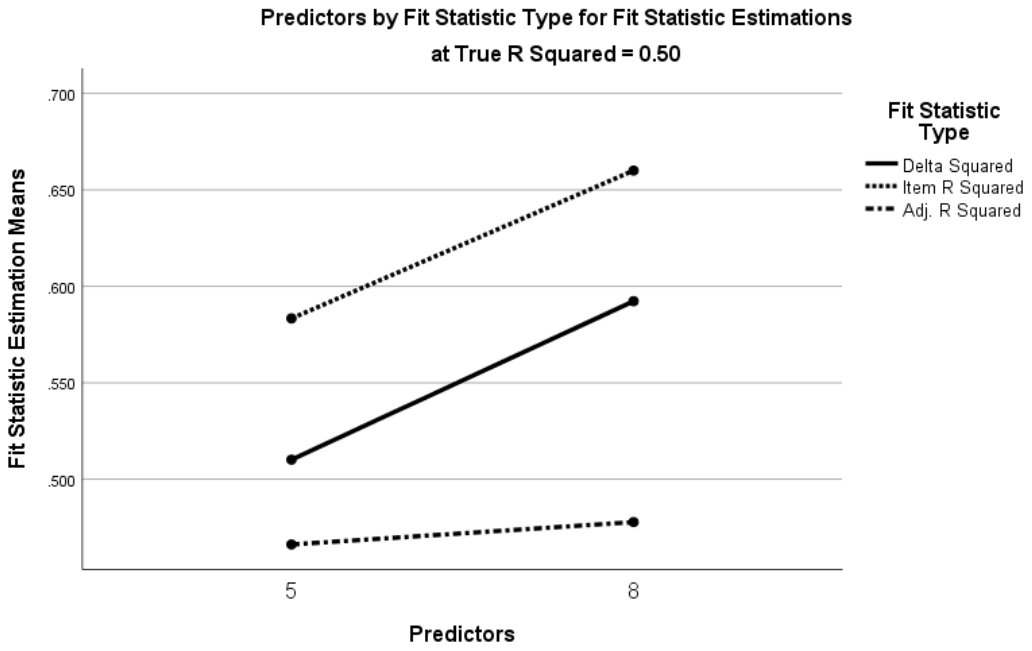

Figure 11 - ANOVA Plot of Predictors by Fit Statistic Type for Fit Statistic Estimations at True R^2 of 0.50

CHAPTER 4. DISCUSSION

The results have several implications. First, the descriptive statistics in Table 1 indicate that in the mean column, which indicates how close the statistic was to the respective true R^2 value, the adjusted R^2 appears to be more favorable because in 6 of 8 cases (except cases 1 and 5 that support Δ^2), the mean was closer to the true R^2 for the adjusted R^2 estimates. However, looking at the standard deviation column of table 1 indicates that the adjusted R^2 also has the highest standard deviation estimates in all cases, which brings the consistency of the adjusted R^2 into question. Looking back at the means, in the cases where Δ^2 was not closest to true R^2, it did have the second closest value to the true R^2, with the item R^2 obviously producing the least desirable means in all cases. Additionally, the adjusted R^2 was only closer than the Δ^2 by 0.05 or less to the true R^2 in 5 out of 8 cases and 0.2 or less in the remaining 3 cases. So, although the adjusted R^2 initially appears to be more supporting, keeping these factors in mind lends much more support to the Δ^2 since its standard deviation values are much lower and the mean estimates are only minimally further than the adjusted R^2 mean estimates.

Second, the results support Δ^2 as having the most optimal measures of mean fit in terms of RMSE and absolute deviation listed in Table 2. Bias calculations initially support the adjusted R^2 as a measure of accuracy; however, looking at individual values and trends of the adjusted R^2 indicated a severe negative trend that may influence bias. Absolute deviation calculations confirmed this suspicion when results supported Δ^2 as a better measure of accuracy over the adjusted R^2, lending full support to Δ^2 as the most optimal measure of both consistency and accuracy. The ANOVA plots corresponding to RMSE

and absolute deviation visually support these estimations as well, showing Δ^2 as being most optimally positioned in more cases than the other two indices.

Third, the results of the ANOVA tables support the significant interactions of most predictors when looking at the measures of fit. In other words, the predictors (test length, number of predictors, and size of true R^2) influence each other in nearly all instances. The four-way ANOVA looking at the estimations were significant up to the two-way interactions (aside from the true R^2 x predictors) as well, indicating a decent amount of influence within variables as well.

It should also be noted the condition assessing 20 items, 8 parameters, and a true R^2 of 0.36 may not optimal. This condition acted differently than the other 7 conditions in that it tended to be the only condition that did not support the Δ^2 when the other conditions did. For example, it was the only condition where the Δ^2 did not have the lowest RMSE and also the only condition where Δ^2 did not have the lowest absolute deviation where, in the other conditions, Δ^2 does have the lowest estimations in both RMSE and absolute deviation. After investigating the individual calculations from the replications, many outliers were found that caused the average of this estimate to become inflated. A possible explanation for this may be that since this condition uses the most extreme conditions for all three predictors, the estimation was more difficult to accurately compute with only 100 replications. A final point of this study is that the number of replications was somewhat of a limitation. While 100 replications provide acceptable results, increasing the number of replications would improve accuracy of the estimations even more and would possibly even reduce the poor estimates of the extreme condition mentioned (i.e. 20 items, 8

predictors, true $R^2 = 0.36$). Future studies will look at increasing replication sizes and the effect of extreme conditions such as this.

CHAPTER 5. SUMMARY AND CONCLUSION

The results from the simulation study are promising support for the Δ^2 index. In most cases, the analyses were favorable toward the Δ^2 statistic, particularly when looking at the comparison of the estimated value to the true R^2 in the item bank, the Δ^2 standard deviation value, and when looking at the ANOVA plots index comparisons. Even when other estimates appeared more favorable, such as the adjusted R^2 showing more estimates closer to the true R^2 or more estimates with lower bias, the adjusted R^2 standard deviation and RMSE scores show the inconsistency of the estimator as compared to the Δ^2 statistic. In comparison, the Δ^2 has shown the most favorable results with both the accuracy and the consistency of the estimates and its error values. Additionally, the earlier prediction made where the Δ^2 statistic and the true R^2 are likely to be closer with more items, fewer predictors, and larger R^2 values was accurate. Table 1 shows the value of the Δ^2 estimate for 30 items, 5 predictors, and an R^2 of 0.5 was 0.503 lending evidence to the accuracy of the Δ^2 model.

The current study examines an alternative to the typical statistical comparisons of nested models when evaluating the quality of an explanatory IRT model, such as illustrated by Dimitrov & Raykov (2003). Similar to structural equation modeling, the index has potential to contribute important information about models beyond statistical tests and information. As illustrated, applications for the LLTM are expanding, but model quality still needs to be assessed. The simulation study provides needed background for an alternative statistic, Δ^2, for evaluating explanatory IRT models.

REFERENCES

Bentler, P. M. & Bonett, D. G. (1980). Significance Tests and Goodness-of-Fit in Analysis of Covariance Structures. *Psychological Bulletin, 88(3),* 588-606.

Dimitrov, D. M., & Raykov, T. (2003). Validation of cognitive structures: A structural equation modeling approach. *Multivariate Behavioral Research*, *38*(1), 1-23.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science*, *50*(3), 328.

Embretson, S. E., & Waxman, M. (1989). Models for processing and individual differences in spatial folding. In *meeting of the Psychonomic Society, St. Louis, MO*.

Embretson, S. E. (1997). Multicomponent Response Models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 305-321). New York, NY: Springer.

Embretson, S. E. (2016). Multicomponent Models. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory, Volume One: Models* (pp. 225-242). Boca Raton, FL: Taylor and Francis Group.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta psychologica*, *37*(6), 359-374.

Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement*, *6*(4), 397-416.

Janssen, R. (2016). Linear Logistic Models. In W. J. van der Linden (Ed.), *Handbook of Item Response Theory, Volume One: Models* (pp. 225-242). Boca Raton, FL: Taylor and Francis Group.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement*, *69*(2), 232-244.

Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement*, *5*(3), 383-397.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (pp. 43-74). Springer, New York, NY.